PhUSE 2015

Paper TT09

CDISC Standards and the Semantic Web

D Iberson-Hurst, Assero, Teignmouth, UK

ABSTRACT

With the arrival of the FDA guidance on electronic submissions, CDISC SHARE and the notion of Research Concepts, the time is ripe to look at improved implementations of the CDISC standards, to assist in producing high-quality clinical research data.

Drawing on experience of production work and the CDISC SHARE project, this paper will examine a prototype implementation to gain insights into the potential of using Research Concepts combined with Semantic Web technologies as the foundation for implementing the CDISC standards, in particular:

1. Review why we want Research Concepts (RCs) and highlight the principles behind them.
2. Look at a prototype semantic web MDR implementation based upon the ISO11179 Metadata Standard, the ISO21090 Healthcare Datatype Standard, the BRIDG model and RCs taken from the CDISC Therapeutic Area development work.
3. Examine prototype tools to see implementation issues and automation opportunities.
4. Detail the benefits that Research Concepts bring to the business and support business artefacts such as annotated CRFs and define.xml.
5. List the existing sources of RC metadata.

INTRODUCTION

How often do we hear the question, “where should I map this?” How many times within a sponsor company does an individual stare at a blank CRF and ponder where to map the various elements? Some mappings are easy to resolve but all too often those faced with this task hit tricky ones. Resources such as the CDISC website, the Implementation Guide (the “IG”), Google, LinkedIn forums, papers from PhUSE and pharmaSUG conferences are searched, looking for a similar situation and example.

Eventually it is decided to map it to X. But in the sponsor company down the road, encountering the same or similar issue, someone maps it to Y or may be X but in a subtly different way. Industry standardization just failed. The FDA picks up the pieces.

We need better standards. There is a need for complete standards that are easy to understand and that do not take a decade to learn and master. They need to be easy to use and, most of all, they need to support the business need.

The work described within this paper is focused on examining whether RCs can address these problems and how to best organize such metadata within a repository of definitions. Also does semantic web technology help us to do so. In simple terms: does this approach help? The best way to answer such a question is to try it out.

RESEARCH CONCEPTS – WHY

VARIABLE-BASED WORLD

Consider a simple example. The Vital Signs (VS) domain is straightforward but serves as an example that people can quickly understand. Within VS we have the test ‘temperature’ as one of a long list of test codes (variable VSTESTCD and the associated code list C66741) and we have a list of units (VSORRESU and the code list C66770) but what is currently not specified is that, when using temperature test, the only valid units are F and C.

Everyone knows this. Unfortunately the machine, the computer, does not. Currently the way we tell the machine is to write specific code, code that is written thousands of times within hundreds of sponsor companies.

Similarly, when dealing with temperature should the test be used in conjunction with VSLOC and/or VSLAT? Does it require VSMETHOD or other variables? With temperature it is easy, again we all know. For more complex tests it is less clear. There is a need to move towards clear and complete definitions in all cases.

RESEARCH CONCEPTS

Research Concepts (RCs) are there to address these issues. Research Concepts simply try to bring all of our knowledge about out data into useful self-contained packages that can be used across the life-cycle. We want to bring clarity to what goes with what, what is valid and what is not. We want complete definitions with ALL terminology defined with the appropriate subsets all in a form that the human AND the machine can understand.

We can then use these definitions within a variety of scenarios as the basis for creating all of our business

objects, supporting the end-to-end process and, very importantly, providing traceability. They can also prove very valuable in impact analysis when, for example, new versions of terminology are produced.

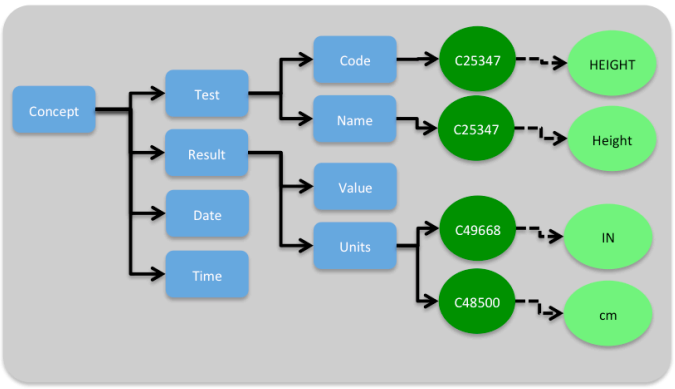
For further background on the end-to-end process please see http://www.assero.co.uk/2015/end-to-end-process-what-is-it/

SIMPLE VITAL SIGNS RESEARCH CONCEPTS

So what do these RCs look like?

Below are two representations of RCs. The figure on the left is a human-readable tree-based view of HEIGHT. To the right is a spreadsheet machine-readable implementation of several VS RCs. The spreadsheet is based upon a simple MS Excel demonstrator implementation used to illustrate the key principles behind Research Concepts. The spreadsheet defines several VS Research Concepts using a simple 'flat' scheme but each can be represented as per the 'tree' diagram seen on the slide. Each Research Concept in this demonstration environment gives:

- Concept Name
- Structure: the tree and the relationships. For example Result is composed of a Value and the Units
- Terminology and subsets, for example Units are 'cm' and 'in'
- Tells us what is not required, no method, no position etc.



	A	B	C	D	E	F
1	**Concept Name**	**Variable Name**	**Code List Ref**	**Code List Item(s)**	**Type**	**Format**
2	Height	Result.Value			I	3
3	Height	Result.Units	C66770	C49668, C48500	E	
4	Height	Test.Code	C66741	C25347	E	
5	Height	Test.Name	C67153	C25347	E	
6	Height	Date			D	
7	Height	Time			T	
8	Weight	Result.Value			F	6.2
9	Weight	Result.Units	C66770	C28252, C48531	E	
10	Weight	Test.Code	C66741	C25208	E	
11	Weight	Test.Name	C67153	C25208	E	
12	Weight	Date			D	
13	Weight	Time			T	
14	HeartRate	Result.Value			I	3
15	HeartRate	Result.Units	C66770	C49673	E	
16	HeartRate	Position	C71148	C62122, C62166, C62167	E	
17	HeartRate	Date			D	
18	HeartRate	Time			T	
19	HeartRate	Test.Code	C66741	C49677	E	
20	HeartRate	Test.Name	C67153	C49677	E	
21	PulseRate	Result.Value			I	3
22	PulseRate	Result.Units	C66770	C49673	E	
23	PulseRate	Position	C71148	C62122, C62166, C62167	E	
24	PulseRate	Date			D	
25	PulseRate	Time			T	
26	PulseRate	Test.Code	C66741	C49677	E	
27	PulseRate	Test.Name	C67153	C49677	E	
28	SYSBP	Result.Value			I	3
29	SYSBP	Result.Units	C66770	C49670	E	
30	SYSBP	Position	C71148	C62122, C62166, C62167	E	
31	SYSBP	Date			D	
32	SYSBP	Time			T	
33	SYSBP	Test.Code	C66741	C25298	E	
34	SYSBP	Test.Name	C67153	C25298	E	
35	DIABP	Result.Value			I	3
36	DIABP	Result.Units	C66770	C49670	E	
37	DIABP	Position	C71148	C62122, C62166, C62167	E	
38	DIABP	Date			D	
39	DIABP	Time			T	
40	DIABP	Test.Code	C66741	C25299	E	
41	DIABP	Test.Name	C67153	C25299	E	

Note: As of August 2015 CDISC decided to rename Research Concepts as Biomedical Concepts. They are synonymous and for the purposes of this paper they will be referred to as Research Concepts.

VITAL SIGNS – ADDITIONAL INFORMATION

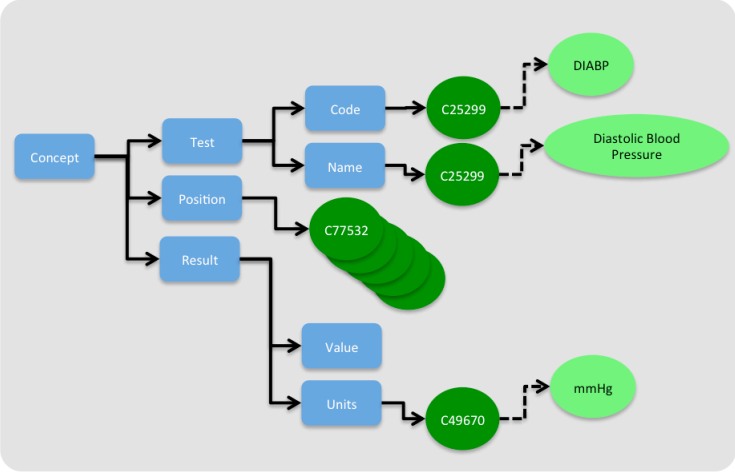
Currently CDISC does not publish RC definitions under an obvious 'RC label'. They do, however, publish information that would be found within such a definition. CDISC released spreadsheets in late 2014 that detailed some of the information that we need for VS and for EG domains.

The first image below details the units for HEIGHT and WEIGHT. Note that for HEIGHT 'mm' is listed as a possible unit whereas it is not included within the simple Research Concept shown above.

	A	B	C	D	E	F	G	H	I
1	C-code (Concept Code)	Vital Signs Test Code (VSTESTCD) (codelist code = C66741)	Vital Signs Test Name (VSTEST) (codelist code = C67153)		C-code (Concept Code)	Units for Vital Signs Results (VSRESU) (codelist code = C66770)		C-code (Concept Code)	Vital Signs Position (POSITION) (codelist code = C71148)
46	C25347	HEIGHT	Height		C49668	cm			
47	C25347	HEIGHT	Height		C48500	IN			
48	C25347	HEIGHT	Height		C28251	mm			
121	C25208	WEIGHT	Weight		C28252	kg			
122	C25208	WEIGHT	Weight		C48155	g			
123	C25208	WEIGHT	Weight		C48531	LB			

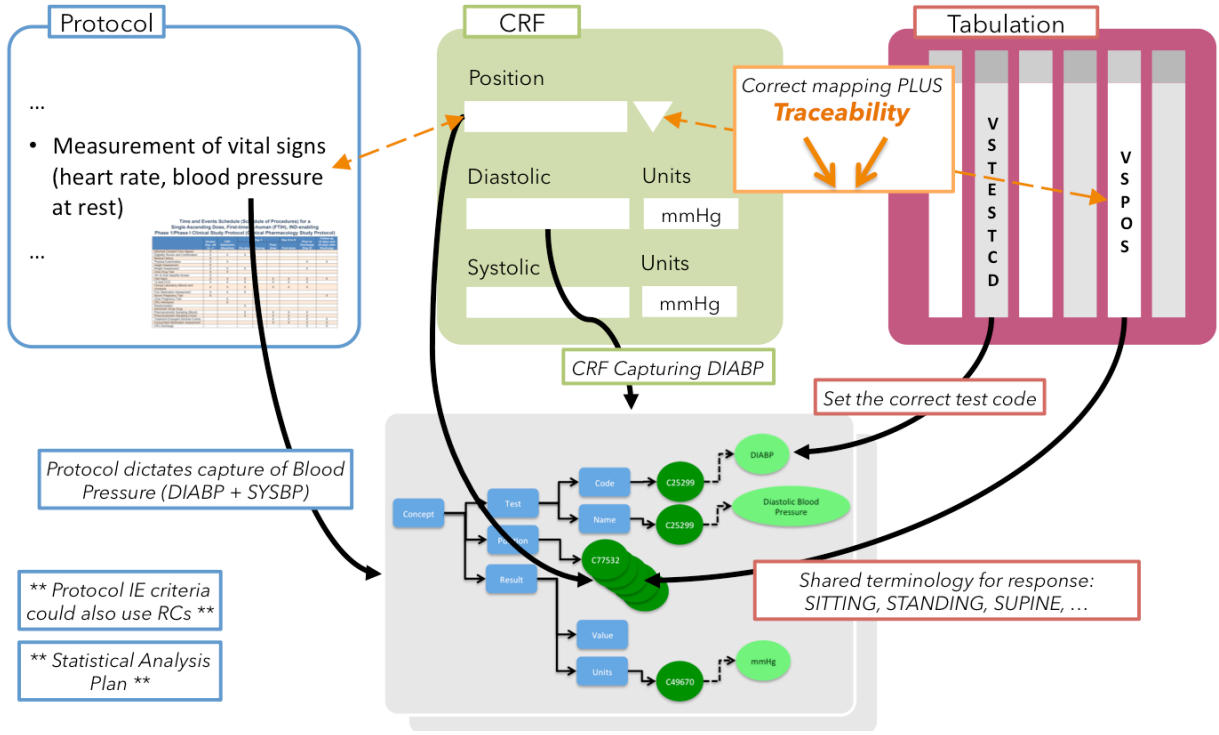
C25298	SYSBP	Systolic Blood Pressure	C49670	mmHg	C62174	SEMI-FOWLERS
C25298	SYSBP	Systolic Blood Pressure	C49670	mmHg	C111310	SEMI-RECUMBENT
C25298	SYSBP	Systolic Blood Pressure	C49670	mmHg	C62122	SITTING
C25298	SYSBP	Systolic Blood Pressure	C49670	mmHg	C92604	SLING
C25298	SYSBP	Systolic Blood Pressure	C49670	mmHg	C62166	STANDING
C25298	SYSBP	Systolic Blood Pressure	C49670	mmHg	C62167	SUPINE
C25298	SYSBP	Systolic Blood Pressure	C49670	mmHg	C62168	TRENDELENBURG

The second image is the equivalent information for SYSBP and DIABP. Note here the additional information relating to position and the code list subset associated with it. Also note that there is a single possible unit, 'mmHg'. This information can be drawn in the human-readable form as above.



Note the structure is essentially the same as the previous HEIGHT RC. This leads to us being able to identify Research Concept Templates (RCTs) that set the pattern for the RCs in a particular area (in this case VS). The templates allow us to build Research Concepts that are consistent and of high quality. RCTs are also aligned with the BRIDG model to help ensure this quality.

DEFINE ONCE, USE MANY
RCs help use reach towards a true end-to-end process.



Looking first at the protocol, often elements like heart rate and blood pressure are mentioned within visit descriptions indicating the tests and/or procedures to be performed within that visit. Here it has been stated that Blood Pressure be captured and that is composed of Systolic and Diastolic pressure.

Is this one concept or two? This question will be ignored for the moment since experience to date is showing that being able to group two Research Concepts into a higher level Research Concept may well be beneficial and thus having BP composed of DIABP and SYSBP would be beneficial.

The use of the Research Concept name within the protocol immediately links the protocol to the Research Concept and can help with subsequent study build processes. It can instantly be seen that collections of, say, Lab Tests (a panel) would be useful as will individual tests or groupings such as questionnaires that relate to several Research Concepts together will provide flexibility to those writing protocols bring clarity on what data are to be collected.

Obviously, also in the protocol is the Visit & Assessment Schedule (Table / Schedule of Events / Assessments / Procedures) that lays out the visit versus assessments for the study. This tends to be more at a CRF level (but is not always, it might be more a collection of RCs) but obviously it could have a direct link to the Research Concepts. It would be desirable to express this table in terms of Research Concepts, Forms built from Research Concepts or groups of Research Concepts.

As an aside two aspects that also are worth considering:
- Linking Research Concepts to inclusion /exclusion criteria to allow for structured protocol IEs that could be machine evaluated.
- The Statistical Analysis Plan and 'analysis concepts' and the use of Research Concepts therein is something that also needs to be investigated, but that is some way off.

The CRF built to capture the data also link to the Research Concept. From the protocol we obviously know which Research Concepts are to be captured and thus each field on the CRF references the corresponding element within the Research Concept which defines the data to be captured along with question text, code lists etc. On this form obviously we need to refer to both the SYSBP and DIABP Research Concepts. Here we have a common field, the position. It will be defined within both concepts but we only want to collect it once, so the form will refer to the same item within both concepts.

Next is the tabulation, the SDTM domain. The VSPOS variable (column) references (points to) the position definitions within the Research Concepts (the other variables similarly pointing to the appropriate item within the Research Concept). This allows the VSTESTCD and VSTEST variables (as well as other fixed values) to be set based on the content within the Research Concept.

A chain has been formed, the protocol specified BP that is composed of SYSBP and DIABP. These Research Concepts are used to define the CRF, the CRF fields referring to the items within the Research Concepts (the leafs of the trees seen earlier) and the variables within the VS domain also refer to the same items. The protocol has been linked to the CRF and onto the tabulation. The links work in the forward direction. Equally, and very importantly, the links can be traversed in the opposite direction; traceability for free!

Silos bedevil our industry. There are standard stacks within each silo but jumping from one to the next is sometimes hard. Much has been done with mapping tools with, for example, define.xml for example with the CRF to SDTM 'gap' but the solutions all leave something to be desired. By using Research Concepts the silos can be knocked down; the CRF can be linked to the Tabulation in both the forward (process) and reverse (traceability) directions. As seen earlier we can also link to the protocol.

Note there are no links into the analysis world as yet but people are thinking on this, there is some work in CDISC and PhUSE but this does need to be attacked. I believe we will see 'concepts' for the analysis work but what form they take it is too early to say.

METHOD

AIM

Given the above the desire was to build metadata based on RCs and prototype business processes to see if RCs will bring benefit while also investigating the benefits the semantic web can bring. To that end a four-step plan was devised as detailed in the rest of this section.

The plan was not the original one. Step 1 was completed with a draft model but there was a need and a desire to quickly visualize the content, it was difficult to keep the entire model as a picture within the brain. Hence, rather than proceeding to step 3 and build a UI for the database immediately, it was decided to implement the model in a simple form and build a quick UI. This allowed "business use" of the metadata to be made in a much shorter timeframe. For more information see http://www.assero.co.uk/2015/a-bit-of-a-tangent/

STEP 1 – A MODEL

Create a semantic model that encompasses all the items needed to meet the business need.

STEP 2 – SIMPLE

Create a simple MDR and Study Build tool to show the ideas working. The tool will use a simple file-based database to speed progress. Use this to assist those grappling with the CDISC content creation.

STEP 3 – SEMANTIC DATABASE

Take the model from step 1 and build a user interface (UI) on top learning the lessons from step 2.

STEP 4 - IMPROVE

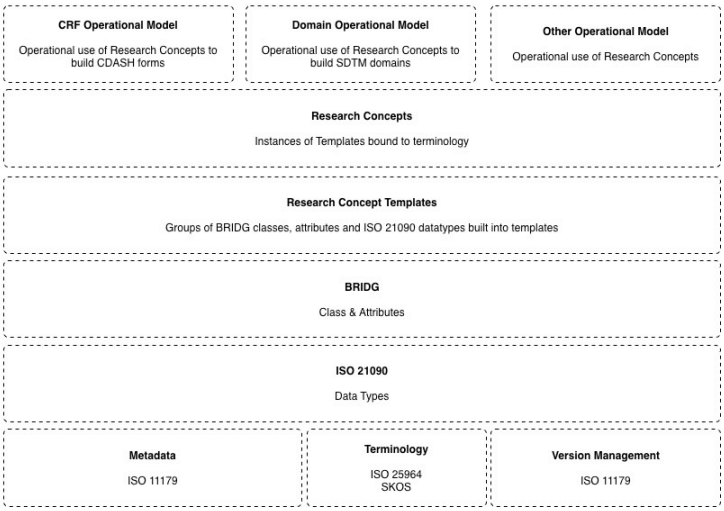
Improve the initial implementation from step 3.

RESULTS

At the time of writing Steps 1 and 2, detailed above, have been completed.

STEP 1 – A MODEL

The output from step 1 was a semantic model. As can be seen from the diagram a layered approach has been taken to build the model:

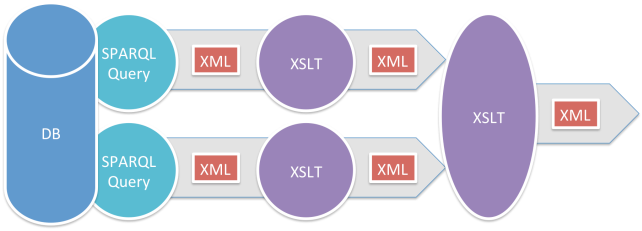


1. Base Layer including Core Metadata Handling (based on ISO11179), Terminology (using ISO 25964) and Version Management (ISO11179 again)
2. ISO 21090 datatypes
3. BRIDG Model
4. Research Concept Templates
5. Research Concepts
6. Operational Models including Forms (CDISC CDASH) and Tabulations (CDSIC SDTM)

It took a while to develop the “standards stack” seen above (see http://www.assero.co.uk/2015/so-many-to-choose-from/ for some background) trying to determine the best approach and which standards to use. Each item within the stack is represented within a semantic model in one or more files with the files were constructed using the Topbraid Composer tool. Once the model was complete sample data was loaded. This included several versions of the CDISC terminology, some RCTs, RCs and then form and domain definitions that used the RCs that had been created. Thus a full model was created and some data populated at each level of that model to allow some simple tests to be run.

The files are available on GitHub and the bottom level has been documented further.

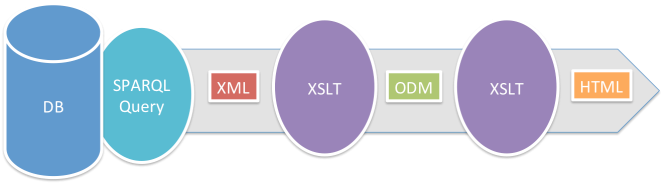
The database was then exercised in two ways. The first was to export the terminology loaded as two versions and to perform a comparison. This was achieved by using a SPARQL query and then some XSLTs to transform and compare the results as shown in the following figure. It was originally hoped to perform the comparison in a query but this was not achieved; this had more to do with a lack of SPARQL knowledge rather than it not being possible.



A second test was undertaken with the database. A single SPARQL query was run against the data to perform an automated annotation of a Case Report Form (CRF). The results of the query where then transformed into an

ODM file and then an old pre-existing XSLT used to converted the ODM file into a human readable HTML form. The process undertaken is represented in the figure below.

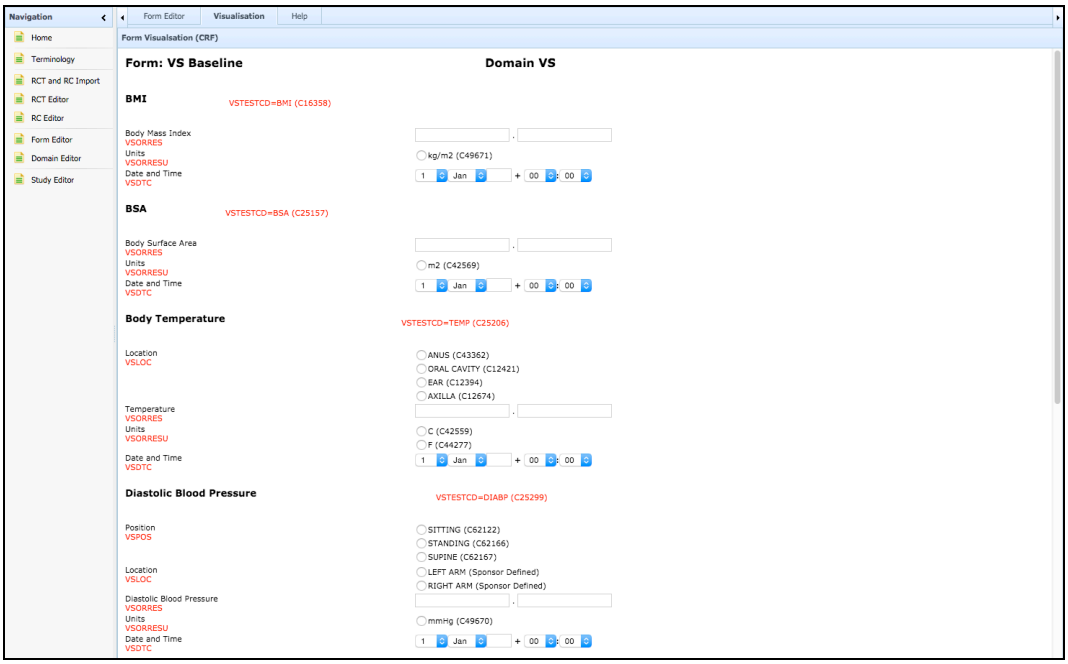
The form used was based on some RCs created by importing some definitions from an existing ODM metadata repository. The “traditional” definitions were passed through an XSLT to convert them into RCs and in a format that could be loaded into the Topbraid tool and the semantic model. These were then incorporated into a simple form definition that was also held within the Topbraid tool.



STEP 2 SIMPLE

The initial step demonstrated the feasibility of using the semantic database but, without a user interface (UI), it was difficult to see some of the benefits. So as to quickly overcome this, it was decided to build a simple UI and a file-based database with that database using the key features found in the semantic version.

The UI included a core set of functions that replicated the business use that a typical sponsor might make of a metadata repository: terminology, creation and maintenance of RC templates, RCs, Forms, Domains and the ability to build a Study. Some of these functions included the ability to visualize the item such as a form. For the form a simple annotated CRF was implemented showing the automation potential of using the RC approach. Below is a screen shot of a CRF that has been automatically annotated.



A further description of the functions developed is available at http://www.assero.co.uk/2015/a-bit-of-a-tangent/ and http://www.assero.co.uk/2015/over-for-another-year/

Another automation opportunity is the creation of define.xml based on a study definition. This is not a difficult task and the work so far indicates that it should be possible to improve implementations of such functionality that have been developed to date. Such a define.xml file could be used to drive study processes such as CRO engagement and data quality checks. It should be noted that, by using Research Concepts, we do not have to expend effort in building value level metadata; it is all present within the Research Concept and we have the complete picture readily available. Thus the creation of define files becomes easier. It should be this way, Research Concepts give us complete metadata which is assembled into a study, define is merely a listing of our study metadata.

One of the aims of this step was to see how much of the complexity of the underlying database could be hidden away from users, who are only really concerned with performing the business functions, while still benefiting from the power that results from the more precise and complete definitions held within the database and the quality and automation that they can bring. The step indicates that this is eminently feasible.

One note to make is that the file database used in this step was again based on CDISC standards, ODM and define.xml files being used extensively for storing RC, Form and Domain definitions.

While undertaking the work an additional requirement arose to demonstrate the ability to link CRFs with healthcare systems (the CDISC E2C work). Quite a bit of work has been done on this topic in the past but having built the tools described above it was possible to examine if using RCs helped with this integration. A small demonstration was prepared whereby two new features where added.

The first function allowed LOINC and UCUM mappings to be added to the terminology section. A small amount of mapping information was entered to allow VS test codes to be mapped to LOINC and the VS units to be mapped to UCUM. The second function was to upgrade the form build function to include metadata that included these mappings.

With these mappings within the form metadata is was possible to write a simple XSLT that took such a form and matched it to data from any CCD document such that the form can be pre-populated. The CCD document is a standard XML document that can be exported by EHR systems within the US.

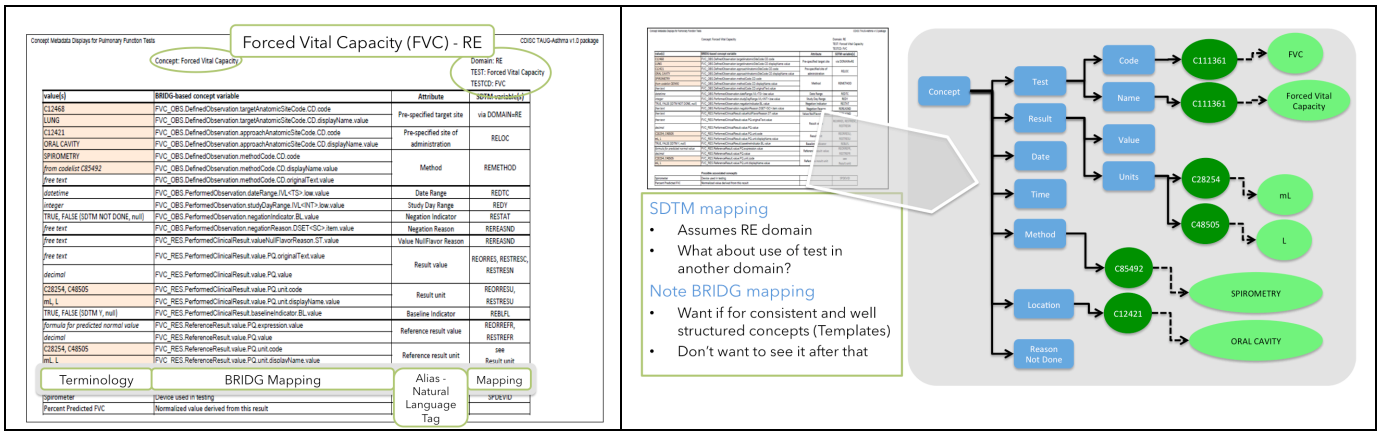
STEP 3 – SEMANTIC DATABASE

The model created in step 1 was implemented locally and run within the Topbraid Composer environment on a laptop, while that built in step 2 was a web-based application. In step 3 the architecture chosen is to split the UI from the semantic database and run both in the cloud.

After a review it was decided to use the Ontotext cloud (the S4 product http://ontotext.com/products/ontotext-s4/) based infrastructure with a UI implemented as a ruby-on-rails implementation running on the Heroku cloud infrastructure. The UI and the Ontotext semantic database communicate via a REST API using SPARQL queries. While it is still early days the results are promising.

Part of the work will be to populate the database with valid CDISC content but where will that content come from given this new RC form? Mentioned earlier CDISC already has some content in spreadsheet form containing partial VS and ECG metadata. That is one source.

The major source will be the Therapeutic Area work currently underway. This diagrams below show an extract from the Asthma Therapeutic Area User Guide (TAUG) and the Forced Vital Capacity (FVC) concept.



Some of the TAUGs, not all, provide very good metadata definitions. If we examine the FVC definition we can extract the metadata and present it as one of the familiar tree diagrams. Note here with this RC that we have two potential result units ("ml", "L") with fixed method and location attributes with a test code of 'FVC'.

CONCLUSION

Research Concepts show great potential. Research Concepts are simply a binding together of all the definitions we need into convenient packages that can be used as the basis of a better, more automated process but that can also be used within existing process to bring significant gains. The industry needs well-defined, well-structured, complete and consistent definitions with all related terminology defined. The use of semantic technology to store those definitions offers benefits from which we can reap great benefit in terms of automation. Looking to the future, if we store the actual data in a similar fashion based upon the definitions even greater gains can be made..

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Dave Iberson-Hurst
Assero Limited,
58 Third Avenue, Teignmouth, Devon, TQ14 9DP, UK
dave.iberson-hurst@assero.co.uk
www.assero.co.uk